Context-based sound event recognition

Niessen, Maria Elisabeth

5

Automatic sound event recognition in the real world

The content of this chapter (except section 5.3) has been published as Niessen, M. E., Krijnders, J. D., & Andringa, T. C. (2009). Understanding a soundscape through its components. In Proceedings of Euronoise.

The content of section 5.3 has been published as part of Krijnders, J. D., Niessen, M. E., & Andringa, T. C. (2010). Sound event recognition through expectancy-based evaluation of signal-driven hypotheses. Pattern Recognition Letters 31(12), 1552–1559.

Human evaluation of sound in real environments is a complex interaction of many factors, which are investigated by a range of research fields. Most approaches to assess and improve the evaluation of sonic environments (soundscapes) use a holistic approach. For example, in environmental psychology, subjective measurements usually involve the judgment of a complete soundscape by people, mostly through questionnaires. In contrast, psychoacoustic measurements try to mimic perceptual attributes by measurements such as loudness. However, these two types of soundscape measurements are aimed at qualitatively different phenomena, which are difficult to link other than with correlational measures. We propose a method grounded in cognitive research to improve our understanding of the link between sound events and human soundscape evaluation. People process sound as meaningful events. Therefore, we developed a model to recognize sound events in a sonic environment that are the basis of these meaningful events.

5.1 INTRODUCTION

Human evaluation of sound in real environments is a complex interaction of many factors. Therefore, many fields of research are involved in trying to understand these factors, ranging from psychoacoustics (Fastl, 1997) to cognitive psychology (Guastavino, 2007) and sociology (Schulte-Fortkamp and Fiebig, 2006). Most psychoacoustic studies on sound quality evaluation focus on measuring attributes such as loudness and sharpness of isolated sounds (Blauert and Jekosch, 1997; Fastl, 1997). However, several studies have shown that these perceptual attributes can only explain part people's evaluation of soundscapes (Ballas, 1993; Maris *et al.*, 2007). The judgment of a soundscape largely depends on the meaning that a person gives to the sound (Dubois *et al.*, 2004). For example, whether a person enjoys music depends on his or her choice to hear it. At a concert, music will be appreciated even (or especially) at a high loudness level, while the tolerance for the music that a neighbor is playing at night will be much lower.

Zhang and Kang (2007) distinguish four categories in which the different factors that influence soundscape evaluation can be organized, namely sound, space, people, and environment. The categories of sound and space comprise the acoustic factors of the sound events in the environment, modified by transmission effects, such as background sounds and reverberation (see chapter 2). These acoustic factors in soundscape evaluation have been tested through psychoacoustic measurements (Genuit and Fiebig, 2006) and questionnaires (Raimbault *et al.*, 2003), amongst others. However, acoustic factors cannot be studied in isolation, because the judgment of people is not based solely on the properties of the sound, but is also affected by the meaning they give to the environment and to the sound events in the environment. This meaning depends on factors like their memory and cultural background. The interplay between acoustic and non-acoustic factors can be examined either by controlling the sounds and varying the condition, such as the cultural background (Hansen and Weber, 2009), or by correlating psychoacoustic measures to people's judgments (Raimbault *et al.*, 2003). Other non-acoustic factors that affect people's judgments can include social, demographical, and behavioral factors (Yu and Kang, 2008), and environmental factors, such as temperature, wind and sunshine (Zhang and Kang, 2007).

Many of these recent studies on the judgment of soundscapes use measure-

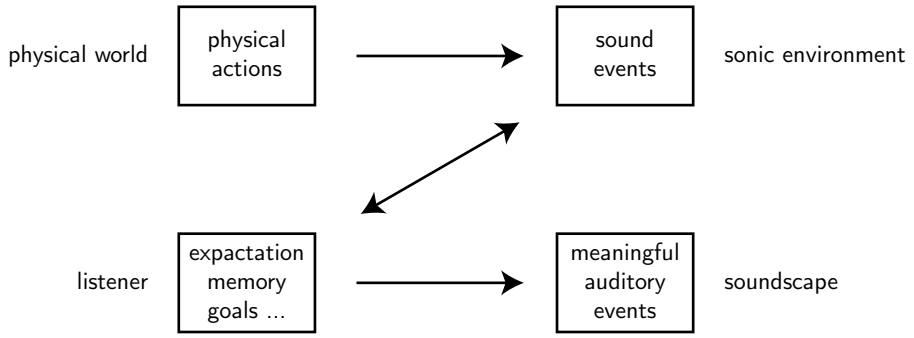

Figure 5.1: Schematic overview of the link between sound events and a soundscape. Physical actions result in sound events, which are interpreted by a listener based on his expectation, goals, and so forth. We refer to a sonic environment as the collection of sound events, and to a soundscape as the collection of meaningful auditory events. (The image is not meant to be exhaustive. For example, the effect of visual information is not included.)

ments that describe the sonic environment as a whole, and not the individual sound events. However, people judge a sonic environment not (only) based on the acoustic properties of the sonic environment. Rather, they evaluate the soundscape by the meaning they give to the environment and the sound events (Guastavino, 2007). For example, Guastavino (2006) showed that an important indicator for the judgment of an urban soundscape is the presence or absence of human activity. Whether there is human activity can be determined by sound events that result from the presence of people, such as speech. Therefore, meaningful events that indicate either a pleasant or an unpleasant situation can function as a link between the holistic judgment of a soundscape and the sound events that constitute the sonic environment (Figure 5.1).

Because sound events are indicators for soundscape evaluation, we propose a method to automatically recognize sound events based on signal-driven hypotheses, which are guided by knowledge of the environment. The recent history of a sound event is used to estimate the context (Box 3.2, page 35) where the event is recorded. Subsequently, the predicted context is used to form expectancies of future events. The hypotheses about events are approximations of meaningful events, because they are learned from human annotations. Furthermore, we use a model

of human memory to represent the hypotheses, through which we include an important cognitive factor in the analysis of a soundscape. Although these event hypotheses are not similar, or even close to human cognitive representations, they can be used to automatically analyze a soundscape based on knowledge other than acoustics. Therefore, this method provides a basis for modeling the factors that are important in soundscape evaluation.

In the next section we will describe the methods that we developed to segregate and interpret sound events. In the third and fourth section we present two experiments with two different data sets to test the integrated methods on their comparison to human annotations. Finally we will discuss the results of the experiments, and provide an outlook on future work.

5.2 METHODS

To recognize sound events in a continuous audio signal, we first segregate patterns from the audio signal that are likely to constitute a single event. Subsequently, we use a model of human memory to select the most likely interpretation for the events that these patterns represent, based on an estimation of the context.

5.2.1 Audio processing

The cochleogram of the audio signal is segmented on the basis of the local spectro-temporal properties. Segments are likely to be produces by a single event when they are based on local properties. For example, local energy maxima that resemble tones and are developing smoothly in time are likely to be produced by the same event. The robustness and reliability of these segments, called signal components, are improved with grouping principles from auditory scene analysis, such as common onset, common offset and common frequency development (Bregman, 1990; Ellis, 1999). Figure 5.2 shows an example of a cochleogram of a speech signal, of which the harmonic components are selected and grouped. The strategy to combine local signal properties and grouping principles allows one to select qualitatively different types of audio patterns, namely tones and harmonic complexes, pulses, and broadband events (see appendix B and Krijnders, 2010). A description of these patterns based on their signal properties is used to classify and label them as sound events with a machine learning algorithm.

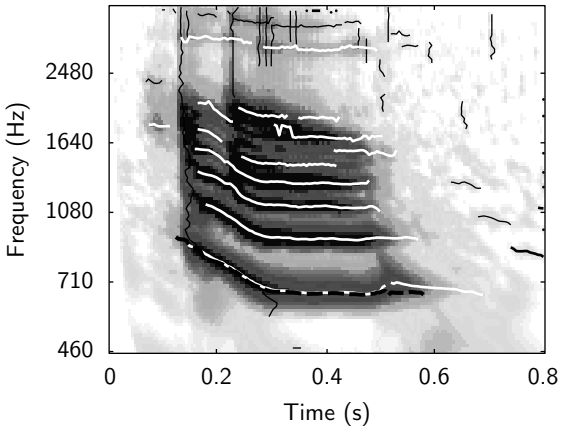

Figure 5.2: Cochleogram of a speech signal with segregated signal components (black lines). The white lines are the components that have been grouped into a harmonic complex, of which the fundamental frequency is depicted by the dashed black line. The vertical black line indicates the onset of the speech.

5.2.2 Context model

The segregated patterns, described in the previous section, are classified with a standard machine learning technique. However, the audio signal can be distorted or masked by transmission effects such as background sounds and reverberation, resulting in a low confidence of certain pattern labels. Furthermore, some sound events can have multiple interpretations while they share similar audio patterns, such as screaming and laughing. To find the most likely interpretation, we introduce a method that incorporates knowledge of the context.

This method, which is inspired by research in cognitive psychology (Quillian, 1968; McClelland and Rumelhart, 1981), constructs a dynamic network that keeps track of both signal-driven patterns and knowledge of the context (see chapter 4). The nodes of the dynamic network represent information about sound events at different description levels, and the vertices between them represent the probability that these pieces of information belong together. Each node holds an activation value. A hypothesis (node) is more likely to represent a relevant interpretation when its activation value is higher than its competitors. Whenever new signal-driven information becomes available, the network is updated by adding nodes,

which each represent new pieces of information, and removing nodes whose activation decreased below a threshold. Subsequently, the activation of the new nodes spreads through the network. Furthermore, new nodes are used to form expectancies of future sound events. If a signal-driven pattern matches an expected event, it is more likely to be an adequate interpretation.

Knowledge network

An example of a network configuration is depicted in Figure 4.3. The nodes at the lowest level represent segregated audio patterns, at the middle level they correspond to sound event hypotheses, and at the highest level to hypotheses about recording locations or sequences of co-occurring sound events. Nodes at the different levels are connected with some strength, denoted by weight w. These weights are learned in the training phase and stored in a knowledge network. In the operation phase the weights are used to infer a probable context of sound events. In the experiments presented in this chapter the context can either refer to a sequence of events (section 5.3) or a recording location (section 5.4).

The strength between the node that represents a recording location or a sequence and the nodes that represent the sound events is calculated according to a term-weighting approach used in automatic document retrieval (Salton and Buckley, 1988). In this method the importance of a term (word or phrase) in a document is determined by multiplying its frequency in the document (term frequency) with its general frequency in other documents (inverse document frequency). Hence, the term is important for a document if it occurs often in that document and infrequently in other documents. Analogously, if a sound event e is encountered often at recording location l, and little at other locations, it is an important indicator for location l. Accordingly, the strength between the sound event e and the recording location l is:

$$w_{l,e} = w_{e,l} = \mathrm{ef} \cdot \frac{\log_{10} N - \log_{10} n}{\log_{10} N} \tag{5.1}$$

where N is the total number of recording locations, n is the number of locations at which e occurs, and the term (that is, event) frequency is given by:

$$\mathrm{tf} = \frac{T_{e,l}}{T_e}, \tag{5.2}$$

where $T_{e,l}$ is the total duration of occurrences of e in l, and T_e is the total duration of occurrences of e in a training set.

The term frequency of events is calculated with the duration of events instead of their frequency, which is common in automatic document retrieval, because duration is more robust to variations in annotations of different human annotators and of different files (see the data descriptions in section 5.3 and 5.4). For example, some annotators choose to annotate every single bird, while others annotate a complete file as containing bird sounds. Therefore, the difference in the number of annotations can be considerable compared to the difference in the duration that is annotated. In other words, what constitutes one instance of an event is more difficult to judge than whether the event is present.

Furthermore, the inverse document frequency (IDF, the second part of equation 5.1) is normalized, such that a sound event that occurs at one location (or is part of one sequence) has an IDF of 1, while a sound event that is recorded at all locations has an IDF of 0, regardless of the frequency it is recorded at those locations (the term frequency). In other words, a sound event that can be heard everywhere does not provide any information about the location of a recording.

Equation 5.1 can also used to determine the weights between sound events and sound sequences, instead of recording locations. If a sound event A is encountered often in combination with some sound event B, and infrequently with other sound events, it is important in the event sequence $s : A-B$.[1] In this case, l in equation 5.1 can be replaced with s, which represents the event sequence instead of the location, so $w_{e,l}$ is substituted with $w_{e,s}$. Additionally, N is the total number of sequences, and n is the number of sequences in which event A of sequence s occurs.

Whether the knowledge network is trained on recording locations or sequences of events can be decided based on the data set. In a data set with recordings at different types of locations, the location can be predictive of the sound event, while this information would be useless in a data set collected at a single location. In such a data set it is more sensible to use location independent information to predict sound events, for example, which sound events co-occur, or at what time they occur.

[1] Two events are considered to comprise a sequence when they co-occur within a certain time frame.

Activation evaluation of event sequences

The spreading of the activation through the network, and the evaluation of the resulting activation values of the hypotheses in the network is determined according to the algorithms that are explained in section 4.3.1. However, the activation evaluation of the event sequences with a fixed order is extended compared to the other hypotheses. Most sequences represent events that can occur in any order. For example, sound events produced by people, such as singing and speech, will generally be heard together, but not in a fixed order. Like all other hypotheses, the expected activation of an event that is part of a non-fixed event sequence is calculated by multiplying the activation value of the event sequence with the connection strength between the sequence and the type of event. Since the activation value decays with time, the expected value is smaller when the other event of the sequence occurred longer ago.

However, for some sequences the order can be very indicative. For instance, in the data set of the first experiment (section 5.3) there are trains departing, which are normally preceded by a whistle of the conductor. Hence, if a whistle is identified, a strong expectancy of a departing train should arise. To capture the regularity of ordered sequences, we determine whether the sound events that constitute a sequence have a strong bias to a specific order. For these ordered sequences, the first sound event primes the network for the second sound event after a time interval learned from examples in the training data. The mean time difference between the events is used in a function to calculate the expected activation value of the second event in the sequence:

$$\hat{A}_i(t) = w_{ji} A_j(t - \Delta t) \mathrm{e}^{\frac{-(\Delta t - \overline{T})^2}{2\sigma^2}}, \tag{5.3}$$

where w_{ji} is the connection strength between event sequence j and expected sound event i, $A_j(t - \Delta t)$ is the previous activation value of event sequence j, Δt is the time interval since j started, and average time interval $\overline{T}$ and standard deviation σ describe the time distribution of the event sequence, as it is learned during the training phase.

Instead of applying a decay to the primed sound event hypothesis (equation 4.3), its activation is determined by weighting the signal-driven evidence with the

expected activation value:

$$A_i(t) = \hat{A}_i(t) + K\left(\frac{n_i(t)}{\max n(t)} - \hat{A}_i(t)\right), \tag{5.4}$$

where $\hat{A}_i(t)$ is the expected activation according to equation 5.3, $n_i(t)$ is the input activation of i as calculated in equation 4.1, $n(t)$ is a list with the input activations of all active sound event hypotheses, and K is the gain factor. The gain factor is dependent on the complexity of the audio signal. If the segregated patterns are very salient, its value should be high. However, sound recorded in real-world environments, as in the presented experiments, are relatively noisy. Therefore, the gain factor in the first experiment is set to 0.5, which entails that the model responds relatively slowly to new evidence from the signal, and is guided equally much by expectancies.

5.3 EXPERIMENT 1

The purpose of this experiment is to demonstrate that the proposed methods can be used to improve the recognition of sound events in a rich outdoor environment, using knowledge of co-occurring events. The data set used in this experiment has been created to develop and test aggression detection systems (Zajdel *et al.*, 2007), and is recorded on a busy train station. Therefore, it includes problems of real-world environments, such as unknown transmission effects and ambiguous sound events. For example, the sound of a train and a subway train are very similar. Based on the audio signal alone, humans have problems recognizing such events as well. In the next section we describe the data set that is used in the experiment. Subsequently, the setup of the experiment is explained, and in the last part we present the results.

5.3.1 Data

The data set consists of 40 enacted scenes from 16 different scenarios, which last between one and two minutes each (Zajdel *et al.*, 2007). The total duration of the recordings is 54 minutes. The scenes were acted by professional actors (three men and one woman) on a platform of the station Amsterdam Amstel. The platform was in normal use by trains on one side and subway trains on the other side. The actors took turns in playing the scenes, such that all scenarios were played out

twice or more with different actors. The 16 scenarios were based on events that are likely to happen at stations, like friends meeting, shouting football supporters, and diverse forms of verbal aggression and vandalism. The scenes were recorded with 8 microphones (16 bits, 44.1 kHz sampling rate), of which one was used for this experiment. This microphone was located about two meters from the center of the action and about two meters from the subway track. The scenes were also captured by three cameras.

The 40 scenes were annotated by the two experimenters who were present at the recordings, based on audio and video, for 13 sound events (Table 5.1). The start and stop times of trains and subway trains, and of some speech, singing and screams, were only indicative, because it was hard to denote the exact times when these events became loud enough to be detectable. Furthermore, to assign a single word to a sound event is often not straightforward. For example, to be able interpret a sound event as either speech or a scream, even given a clear scenario, depends on knowledge of the expressing person and the situation, and one word is usually not sufficient to describe an event. However, we chose to annotate the sound events with one word only. As a result, the performance benchmark for the system is one-dimensional but contentious. In section 5.5 we discuss the considerations for annotation procedures and performance benchmarks further.

5.3.2 Setup

The annotations of sound recordings were used to train both a naive Bayes classifier from the Weka toolbox (Witten and Frank, 2005) for the classification of the signal-driven patterns, and the knowledge of the context model. For the naive Bayes classifier, all 40 audio files (each containing one scene) were processed with the signal-driven method described in section 5.2.1. The segregated harmonic complexes with the highest score given by the harmonic complex grouping algorithm—this score is based on the correspondence of the segregated harmonic complex with an ideal harmonic complex, see appendix B—that overlapped with an annotation, were given that annotation as a label. Harmonic complexes without temporal overlap with an annotation were labeled as indefinite. All other harmonic complexes were discarded. Furthermore, pulses and broadband noises that overlapped with an annotation were labeled. From these processed audio files, 40 file pairs were

Table 5.1: Sound events annotated in the audio data recorded at the Amsterdam Amstel station, their occurrences, duration, and duration as part of the total duration of the data set.

Sound event label	Total occurrences	Total duration	Part of total
Speech	521	6 min 10 sec	16%
Scream	290	3 min 33 sec	10%
Singing	82	2 min 32 sec	7%
Subway	40	5 min 22 sec	18%
Kick	26	9 sec	0.4%
Train	15	3 min 20 sec	9%
Subway door signal	14	18 sec	1%
Laughing	12	13 sec	1%
Train whistle	3	5 sec	0.2%
Subway horn	3	3 sec	0.1%
Announcement signal	2	4 sec	0.2%
Birds	2	3 sec	0.1%
Train door signal	1	3 sec	0.1%

generated that contained the feature vectors describing segregated patterns. Each file pair consisted of a file used for training, for which the feature vectors of the segregated patterns from 39 files were used, and a test file, which contained the feature vectors of the segregated patterns from the one file that was left out, resulting in a leave-one-out set.

Additionally, the annotations of the training file of each file pair were used to train the weights in the dynamic network (section 5.2.2). On average 18 different types of sequences were encountered in the training set. These sequences are composed of the 13 sound events listed in Table 5.1. An average of 89 examples of each sequence was used to train the weights in the knowledge network. The spread of the number of examples per sequence was very large, ranging from 2 to 730.

In the test phase, the patterns in the test file are classified with the naive Bayes classifier and used as input for the dynamic network ($\tau = 100$ in equation 4.2, activation threshold $\theta_A = 0.2$, and $K = 0.5$ in equation 5.4). Subsequently, the possible sound events that the pattern can represent are initiated as hypotheses in the net-

work. The weight between the pattern and the event hypotheses is the probability of each event label given by the naive Bayes classifier. If the event cannot be classified and is labeled as indefinite, the weight is set to the default activation of the event. Based on these events, the network forms a hypothesis of a sound sequence, which in turn initiates expectancies of certain sound events that might follow. The results of this integrated approach are the mostly likely events that explain the segregated patterns, given the recognized sound events and their recent history.

The most likely events according to the naive Bayes classifier and the integrated model are compared to the human annotations through the *F*-measure. The *F*-measure is used in information retrieval to test the effectiveness of the performance of a system (Van Rijsbergen, 1979), for example a search engine. The *F*-measure is computed as the harmonic mean between the recall, which represents whether relevant results are retrieved, and the precision, which represents whether irrelevant results are not retrieved.[1] Applied to the results of automatic sound recognition in our experimental setup, precision is a measure for the fraction of time the recognitions are correct, and recall is a measure for the fraction of recognitions that are made out of the amount that should have made compared to the annotations.

$$\begin{aligned} \text{precision} &= \frac{TP}{TP+FP} \\ \text{recall} &= \frac{TP}{TP+FN} \\ F &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \end{aligned} \tag{5.5}$$

where TP is the true positive rate, FP is the false positive rate, and FN is the false negative rate.

5.3.3 Results

The event sequence prediction is based on the classified segregated patterns, and used to select the most likely interpretation for the pattern. Of all 13 types of annotated sound events, 7 are recognized (segregated and labeled) by the Bayes classifier and the integrated model (the segregation algorithm, the Bayes classifier, and

[1] The harmonic mean tends strongly toward the least of the two values. Hence, it penalizes a focus on only one of the two measures.

Table 5.2: *F*-measure, precision, and recall of random classification (R), a Bayes classifier (C), and the integrated model (I) per sound event type.

	F-measure			Precision			Recall		
Sound event label	R	C	I	R	C	I	R	C	I
Singing	0.07	0.28	0.48	0.07	0.26	0.36	0.07	0.31	0.72
Speech	0.16	0.18	0.38	0.16	0.50	0.44	0.15	0.11	0.33
Train	0.09	0.40	0.43	0.09	0.32	0.35	0.09	0.54	0.55
Subway door signal	0.01	0.28	0.26	0.01	0.25	0.21	0.01	0.33	0.33
Subway	0.17	0.52	0.57	0.18	0.54	0.62	0.17	0.49	0.53
Kick	0	0.28	0.24	0	0.26	0.65	0	0.30	0.14
Scream	0.09	0.36	0.44	0.10	0.36	0.36	0.09	0.36	0.56

the context model). These 7 types of sound events were most frequently annotated (see Table 5.1). Hence, the Bayes classifier and the context model can learn them more reliably than events that occur infrequently. Table 5.2 shows the *F*-measure, precision and recall of the recognitions made by the Bayes classifier (C) and by the integrated model (I) for the 7 sound event types. These measures can be compared to a random classification (R), which is based on the amount of time that the sound events are annotated, that is, the a priori probability that the sound events are encountered. Figure 5.3 displays the overall results of both models.

On average, the context model improves the *F*-measure compared to the signal-driven classification, mostly through an increased recall of the sound events that have more harmonic content (singing, screams and speech), because these types of events are more likely to be of the same type as their surrounding events. For example, the network may change a speech classification to a scream when surrounded by screams. If this change is correct, both the recall of the scream event and the precision of the speech event increase. However, the increase in precision of harmonic sound events is moderated by some erroneous changes in other sound events. As a result, the overall precision does not increase as much as the recall.

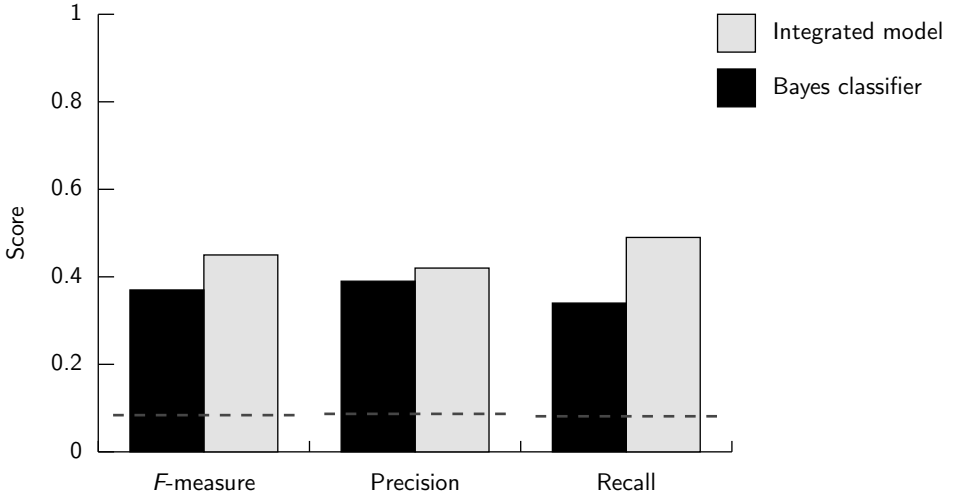

Figure 5.3: Overall results of the Bayes classifier and the context model integrated with the Bayes classifier on the data set recorded at the Amsterdam Amstel train station. The dashed lines show the overall performance of random classification, which is based on the average amount of time that the sound events are annotated.

5.4 EXPERIMENT 2

This experiment demonstrates that the proposed methods can be used to recognize sound events in a sonic environment given a predicted location. Furthermore, both the environment and the recorded events are uncontrolled in the data set of this experiment. As a consequence, the diversity of the recorded events is enhanced compared to the first experiment. In the next section we describe the data set that is used in the experiment. Subsequently, the setup of the experiment is explained, and in the last part we present the results.

5.4.1 Data

The data was collected under different weather conditions on a number of days in March 2009 in the town of Assen (65,000 inhabitants, in the north of the Netherlands). The recordings were made by six groups of three students as part of a master course on sound recognition. Each group made three minute recordings at six different locations: a railway station platform, a pedestrian crossing with traffic

Table 5.3: Examples of sound events annotated in the data set recorded in Assen, their occurrences, duration, and duration as part of the total duration of the data set.

Sound event label	Total occurrences	Total duration	Part of total
Bird	238	17 min 6 sec	11%
Bike	30	2 min 22 sec	2%
Rooster	16	43 sec	0.5%
Horn	8	11 sec	0.1%
Shopping bag	1	7 sec	<0.1%

lights, a small park-like square, a pedestrian shopping area, the edge of a forest near a cemetery, and a walk between two of the positions. Recordings were made using M-Audio Microtrack-II recorders with the supplied stereo microphone at 48 kHz and 24 bits stereo.

All the recordings were annotated by two students separately. The average agreement between the two annotators of one group was determined with the F-measure (equation 5.5): $\overline{F} = 0.46$ with a standard deviation of 0.25. These two annotations were merged, such that equal labels did not overlap, but became one instance (a union in set theory). We examined the resulting merged annotations, and adjusted them when necessary. However, we did not introduce new annotations.[1] We ensured that the names of events were uniform across all the files to prevent the context model from learning annotators rather than locations. The total of 44 audio files, with an average duration of 3.5 minutes, were annotated for 54 different sound events. However, half of these sound events were annotated less than 5 times, while just a few events comprised most of the annotations. A few examples of annotated events are given in Table 5.3, ranked according to their frequency in the complete data set.

5.4.2 Setup

The experiment was designed in the same manner as the first experiment (section 5.3.2). We used a nearest-neighbor classifier instead of the naive Bayes classifier

[1] An exception was made for the annotations of one group, which we had to complete because they were sloppy or omitted.

to label the segregated patterns. For the nearest-neighbor classifier, all 44 audio files were processed with the signal-driven method described in section 5.2.1. The segregated harmonic complexes with the highest score given by the harmonic complex grouping algorithm—this score is based on the correspondence of the segregated harmonic complex with an ideal harmonic complex—that overlapped with an annotation were given that annotation as a label. Harmonic complexes without temporal overlap with an annotation were labeled as indefinite. All other harmonic complexes were discarded. Furthermore, pulses and broadband noises that overlapped with an annotation were labeled. From these processed audio files, 44 descriptive file pairs were generated that contained the feature vectors describing segregated patterns. Each file pair consisted of a file used for training, for which the feature vectors of the segregated patterns from 43 files were used, and a test file, which contained the feature vectors of the segregated patterns from the one file that was left out, resulting in a leave-one-out set.

Additionally, the annotations of the training file of each file pair were used to train the weights in the context model (section 5.2.2). The information used to train the weights between the locations and the sound events is summarized in Table 5.4. Furthermore, it shows a few examples of the connection strength between the locations and some sound events. On average 21 sound events were encountered at each recording location in the training set. Furthermore, an average duration of 80 seconds per sound event at a location was used to train the weights in the knowledge network. The spread of duration per sound event was very large, as can be seen in Table 5.3.

In the test phase, the segregated patterns in the test file, which are classified by the nearest-neighbor classifier, are used as input for the dynamic network ($\tau = 100$ in equation 4.2 and the activation threshold $\theta_A = 0.2$). Subsequently, the possible sound events that the pattern can represent according to the learned knowledge are initiated as hypotheses in the network. The weights between the pattern and the event hypotheses are the probabilities of each event given by the nearest-neighbor classifier. If the event cannot be classified and is labeled as indefinite, the weight is set to the default activation of the event. Based on these event hypotheses, the network forms a hypothesis of the location, which in turn initiates expectancies of certain sound events that might follow. The results of this integrated approach are the mostly likely sound events that explain the segregated patterns, given the recog-

Table 5.4: Locations. $\overline{N}$ is the average number of sound events annotated at a location, $\overline{T}$ is the average of the total duration per annotated sound event at a location in seconds, and $\overline{s_T}$ is the average spread of the event durations at a location. The two weights $w_{l,\text{train}}$ and $w_{l,\text{bus}}$ are examples of learned connection strengths between locations (l) and sound events.

Location	$\overline{N}$	$\overline{T}(s)$	$\overline{s_T}$	$w_{l,\text{train}}$	$w_{l,\text{bus}}$
City center	23	64	152	0	0
Graveyard	15	100	157	0	0
Museum	24	68	93	0	0.02
Traffic lights	17	87	181	0	0.15
Train station	26	79	111	1	0.02
Walking	21	81	170	0	0.04

nized sound events and their predicted location. The most likely events according to both the nearest-neighbor classifier and the integrated model are compared to the annotations through the F-measure (equation 5.5).

5.4.3 Results

The success of the context model as it is applied in this study is dependent on whether the recording location prediction is correct. The results of the location predictions of the test files are listed in Table 5.5. The number of test files at each location is in parentheses behind the location name. The location predictions of the 7 test files of recordings during walking are not included, because they cannot be assigned to a single location. The top 1 indicates the amount of time the location predictions are correct on average for a specific location. The model has an activation or confidence value for all the location hypotheses. Therefore, if the best prediction is not correct, the second best might be. The top 2 and 3 specify whether the correct location is among the second or third best predictions.

Only two locations can be predicted well, the train station and the museum, because the some sound events that are specific for those locations, such as train sounds for the train station, are segregated and correctly classified. In contrast, many of the other sounds that can be segregated and classified by the nearest-neighbor algorithm, such as cars and speech, are generic, and can be heard at any of

Table 5.5: Results of location predictions: the average amount of time that the location prediction was correct per location. The number of files recorded at each location is shown in parentheses.

Location	Top 1	Top 2	Top 3
City center (7)	1%	2%	2%
Graveyard (7)	1%	1%	15%
Museum (8)	6%	87%	93%
Traffic lights (7)	2%	8%	59%
Train station (8)	90%	98%	98%

the locations. Therefore, the location prediction is not reliable in many test files.

The location prediction is based on the classified segregated patterns, and used to select the most likely label for the pattern. Of all 54 annotated sound events, 11 events are recognized (segregated and labeled) by the nearest-neighbor classifier and the integrated model (the segregation algorithm, the nearest-neighbor classifier, and the context model). These 11 sound events are most frequently annotated. Hence, the nearest-neighbor classifier and the context model can learn them more reliably than sound events that occur infrequently. Table 5.6 shows the *F*-measure, precision and recall of the identifications made by the nearest-neighbor classifier (C) and by the integrated model (I) for the 11 sound events. These measures can be compared to a random classification (R), which is based on the amount of time that the sound events are annotated, that is, the a priori probability that the sound events are encountered. Therefore, the results of the random classification are relatively high on sound events that are annotated often and long, such as speech and footsteps.

On average the context model slightly improves the *F*-measure compared to the signal-driven classification, mostly through an increased recall, which means that more correct instances of annotations are found than with the nearest-neighbor classifier (Figure 5.4). For both models the performance on this data set is lower than the performance on the data set recorded at the Amsterdam Amstel train station, because this data set is more divers, and recorded in less controlled conditions. Furthermore, the first data set is annotated by two people, compared to 14 in this data set, resulting in more diverse annotations.

Table 5.6: *F*-measure, precision, and recall of random classification (R), a nearest-neighbor classifier (C), and the integrated model (I) per sound event.

	F-measure			Precision			Recall		
Sound event label	R	C	I	R	C	I	R	C	I
Bird	0.109	0.005	0.053	0.112	0.093	0.198	0.106	0.003	0.031
Braking train	0.018	0.193	0.194	0.019	0.274	0.188	0.018	0.149	0.201
Bus	0.026	0.053	0.055	0.026	0.201	0.029	0.026	0.031	0.423
Car	0.195	0.499	0.408	0.203	0.594	0.427	0.184	0.431	0.391
Footsteps	0.173	0.008	0.104	0.181	0.091	0.358	0.166	0.004	0.061
Passing train	0.005	0.570	0.612	0.005	0.416	0.463	0.005	0.906	0.906
Rain drops	0.044	0	0.052	0.045	0	0.135	0.044	0	0.033
Speech	0.127	0.023	0.111	0.131	0.223	0.124	0.123	0.012	0.100
Starting train	0.019	0.022	0	0.019	0.018	0	0.019	0.028	0
Truck	0.057	0.021	0.013	0.057	0.045	0.010	0.056	0.014	0.017
Wind	0.129	0.071	0.138	0.134	0.134	0.135	0.125	0.048	0.141

5.5 CONCLUSION

In the previous sections we have demonstrated that a model that combines both signal-driven algorithms and contextual knowledge in the form of predicted recording locations or event sequences, improves the recognition of sound events in a real-world environment compared to an exclusively signal-driven method. Especially the events that have similar audio patterns, and hence rely more on context for their interpretation, for example screams, speech and singing in the first experiment, are recognized better in the integrated approach. Sound events that are already recognized well by the signal-driven algorithm, such as trains and cars, gain little improvement from the context model. Finally, sound events that occur infrequently, and hence have few training examples, and events that are not yet captured well by the audio features vectors, show a small performance reduction.

The overall results (especially of the second experiment) may not seem impressive, but this is partly explained by the performance measure. The *F*-measure is based on the temporal overlap of the annotations and the labeled patterns. Therefore, it is dependent on both the annotations and the segregation algorithm. Annotating sound is a complex process, which is demonstrated by the low inter-

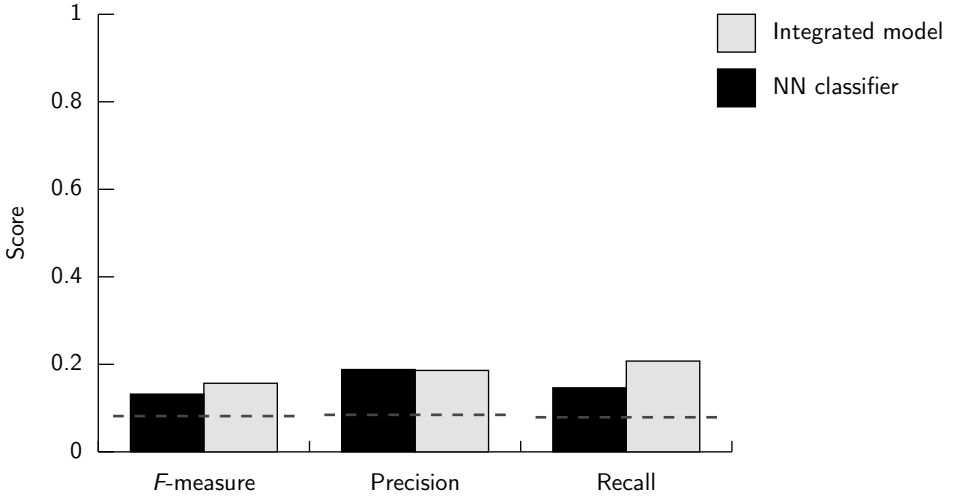

Figure 5.4: Overall results of the nearest-neighbor (NN) classifier and the context model integrated with the nearest-neighbor classifier on the data set recorded at different locations in Assen. The lower dashed lines show the overall performance of random classification, which is based on the average amount of time that the sound events are annotated.

annotator agreement ($\overline{F} = 0.46$). The annotators did not only use information in the sound, but also knowledge of the environment, because they were present during the recordings. We cannot determine to what extend the annotations are based on the sound or on their knowledge. Some annotated sound events can even hardly be recognized by a human who has to rely on the audio signal alone.

In contrast, the segregation algorithm relies solely on the audio signal. This signal is uncontrolled and thus very challenging for the algorithm that needs to segregates relevant parts. For example, people annotate sound events that occur in a sequence, such as footsteps and birds, continuously even though they are interrupted.[1] In contrast, the segregation algorithm has to segregate every single occurrence before it can be recognized. As a consequence, the precision of both models on these types of sound events can never be high. Furthermore, especially the recordings in the second experiment contain a wide variety of sounds events,

[1] The random classification outperforms the models on a few of these sound events, because it is based on a priori probability of occurrence determined from the annotations.

most of which occur only a few times in all the recordings.[1] To be able to learn the patterns of a sound event, a machine learning algorithm (like Bayes or nearest-neighbor) needs more examples than were available of most sound events in the data set in these experiments.

These observations demonstrate that modeling context information is essential to achieve robust event recognition in real-world environments. Indeed we have shown that context, in the form of recent history of sound events, improves sound event recognition, even though it is so far only based on knowledge derived from the human annotations of the audio signal. Additionally, the recording location is not predictive for many generic sound events, such as speech and cars[2], which occur most often, and are classified best by the nearest-neighbor classifier. In other words, the infrequent events are the events that are good predictors of a location, while these are the hardest events to learn, because they are infrequent. However, the context model is not limited to process acoustic information. In the next chapter we show that the context model can also be used to process ambiguous visual information. Because the model can receive input from different modalities, it can combine multiple modalities and factors in a single system that returns a single analysis. We plan to integrate information from multiple sources of knowledge so that the context is modeled more profoundly.

In summary, the integrated model provides a new strategy to analyze sonic environments by recognizing sound events. Because these sound events are also based on knowledge of the context, they are a first approximation of meaningful auditory events. However, to improve the recognition of these events in the complexity of real environments, we require a combined development of segregation algorithms and models that can include non-acoustic factors. Furthermore, we will study human perception in parallel, so we can validate the model for soundscape analysis. Conversely, the development of a system to analyze a soundscape automatically might increase our understanding of human soundscape evaluation.

---

[1] The average amount of time that each sound event is annotated is 2%. Excluding the few sound events that occur most often (birds, cars, footsteps, speech, and wind), this amount is less than 1%.

[2] In the knowledge network of the second experiment, the weights between speech and cars and all locations were 0.